J. Appl. Environ. Biol. Sci., 4(7S)113-120, 2014

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Automated Career Counseling System for Students using CBR and J48

Maha Nawaz, Anum Adnan, Unsa Tariq, Jannat Fatima Salman, Rabia Asjad, and Maria Tamoor

Department of Computer Sciences Kinnaird College for Women Lahore, Pakistan

Received: September 1, 2014
Accepted: November 13, 2014

ABSTRACT

Career-related confusions are a serious issue among students these days. A student needs an accessible, easy-to-relate-to and trustworthy career planning resource at their disposal. Mostly at age of 18, students are not mature enough to know precisely what career to follow; they are not sufficiently aware of what goes on in a particular area and which academic majors are associated with their areas of interests. This paper presents an automated system that mimics a one-to-one meeting with a professional career counselor. The system supports people in developing their own career opting competences. The paper focuses on collating different machine learning algorithms to guide students on the basis of their academic background, hobbies and location. This proposed system helps student in their University choice, career choice and the scope of their respective appropriate career for future.

KEYWORDS— CBR, machine learning algorithms, career planning, student, A levels

1. INTRODUCTION

The selection of career paths for students after A-levels/intermediate is an attention requiring concern. A recent survey of 115,000 people from 33 different countries indicated that 50% of the people felt that they had chosen the wrong career ("Mail's Globe Careers", 2009). Notably in Pakistan, there is not enough guidance for students, either from their institutes or other sources, which can instruct them to adopt such majors in University that is best suited to their interests and skills. Present paper confronts the career-related confusions amongst students nowadays. This research intends to solve the career assortment problems by making use of the CBR (Case Based Reasoning) and Decision Tree J 48 algorithm. The system establishes an automated process similar to a one-to-one meeting with a career counselor and aids to 'plan' a career true to the student's grade, IQ, hobbies and, predominantly, gender. Students can later determine a career from the proposed options and the illustration of related jobs. The system's distinction is to nominate Universities offering education for the recommended careers.

Rest of the paper is organized as: Section 2 includes the related work done concerning student's educational and career decision problems. Section 3 describes the methods involved in the building and programming of the system. Section 4 discusses the data collected, experiments done for the system and its results. Section 5 compares the techniques employed. Sections 6, 7 and 8 conclude the paper.

2. RELATED WORK

Career counseling is based on a student's previous academic performance, skills and potential and students are often unaware of what may suit these attributes perfectly. In that manner, career counseling is similar to defining students' problems related to learning and application and their solutions. Many papers are written that propose solutions to students' related problems by implementing data mining techniques along with certain machine learning algorithms. The overall goal of the data mining process is to extort information from a data set and transform it into a rational structure for further use. Misinterpretation related to career counseling is the major problem faced by undergraduate students. For this purpose, intense examination of certain input is important for effective student development through effective career counseling (Hall, 2005). Multiple researches have been conducted to acquire student's educational attributes to observe future career patterns. With the help of certain algorithms, career related decisions are deduced. The inference of these researches leads to career guidance based on their transitional period. Pal et al. (2014) propose data mining techniques for identifying patterns in vast databases of multiple universities to investigate alumni and students' challenges regarding career and counseling. Kakavand et al. (2014) applied Decision Tree algorithm to process post-graduate students' academic information and predict the attributes of those students who are inclined to pursue their studies on the basis of the pattern identified from a database of Post-

*** Corresponding Author:** Maha Nawaz, Department of Computer Sciences Kinnaird College for Women Lahore, Pakistan.
Email: maha_nawaz02@yahoo.com,

Graduate students. Cao et al. 2012, elaborated college students' complications regarding their present career choices. In order to aid these students in determining their professional problems they utilized basic career counseling, information and evaluation and management along with auxiliary decision-making. Thus, proposing solutions and recommendations by assessing effectively and employing agent technology to create a web based system. (Stebleton, 2006) discussed the theoretical approaches of counseling African students in the US Universities together with its practical implications. Yadav et al. (2012) applied ID3 algorithm, C 4.5 algorithm and CART algorithm on a student database to predict Engineering student's performance in final exams. Baradwaj et al. (2011) resort to data mining techniques in order to improve quality of higher education. Conati et al. (1997) have worked on an online model for coached problem solving by operating Bayesian networks and incorporating ANDES, which is an intelligent Tutoring System for Newtonian physics. They used stochastic sampling algorithms to update the network and predict students' actions during problem solving. Hasebrook et al. (1997) lodged an expert advisor that works on many platforms that gives vocational guidance by making use of expert advice for the same input. The limitations with the system are that it does not cater alien inputs and provides career suggestion for only renowned majors. (Miller, 2006) proposed a solution-focused counseling strategy for career counselors to better advice careers to clients who seek only a little direction rather than letting the counselor control their professional choices. It enables career counseling practitioners to induce self-helpfulness in such clients. The outcome from the application of such strategies is mostly helpful when clients come for counseling for only one session. Schedin (2007) sparsely investigated the interaction process between a client and career counselor to describe and analyze interpersonal behavior in career counseling sessions. This research was driven by interpersonal theory and the model of structural analysis of social behavior (SASB) developed by Benjamin S., Feduccia (2003) explained the influence of Career Discovery I, which is the first module in a computer-assisted program for making career decisions, on the firmness of choice of college majors. The research determined differences, if any, between students who entered a Research-extensive University without declaring a major to those who declared one. Crozier et al. (1985)focus on the role and function of post-secondary career counseling specialists and the issues that affect their practical implementation within post-secondary institutions. Brown et al. (2002) suggest how social cognitive career theory's major hypotheses can be applied to counseling careers and to develop a broad array of career choices and analyzing its barriers as well as how to overcome those barriers. Betz et al. (2014) review the literature on Bandura's (1997, 1982) self-efficacy theory to the career field and describe the usefulness of career self-efficacy in building such models that predict the occupational choice behavior of men and women along with understanding the disadvantaged status of women in labor force through elf-efficacy utility. Ye (2014)] applied the same for Chinese Graduate students. [18] describes the demographic, economic and social effects on career preparations. (Suling, 2012) made use of multi-objective decision-making and combines it with BP artificial neural network appropriately to construct the general diathesis estimation model for university students. (Westbrook B, 1999) is the replication of past studies by using an improved design to determine the relationship between aptness of career choices and career maturity test scores.

However, the entire above papers proved that no work has yet provided with an optimized and accurate output in regards to students' problems with career choices after A-levels/Intermediate. But in this paper, automated career planning and data mining techniques are integrated with CBR and Decision Tree J48 to distinguish the academic, personal and intellectual patterns of A levels/intermediate students for productive career guidance.

3. METHODS

Automated Career Planning incorporates concepts of Artificial Intelligence and Machine Learning Algorithm which proposes a suggestion to students regarding majors most appropriate for them. It is a formulated technique for analyzing an individual's abilities through his/her interests and hobbies. Thus constructing solutions for students related to career planning problems.

The system's fabrication involves more than one algorithm from Weka, a collection of Machine Learning algorithms for data mining tasks also the algorithm involved CBR of hamming distance with Manhattan distance to calculate the output as well as to refine the accuracy of the results.

System is designed in a way that it takes inputs from the user, matches it with the training data and yields an output. Following are the fields that the user fills as inputs:

1. Name(String)
2. Gender(Char-->F/M)
3. High School Grade(Char--->A-F)
4. Hobbies (Radio Buttons)
5. Skills (Radio Buttons)

6. IQ Grade(Char--->A-F)

3.1 Case Based Reasoning (CBR)

In case-based reasoning (CBR), career counseling system's capability is exemplified in an archive of past cases, instead of being encoded in traditional rules. Each case typically comprises an explanation of the problem, along with an answer and/or the output or result. The knowledge and reasoning process demonstrated by an expert to solve the case is not noted, but is contained in the solution.

To solve a unique case, it is matched against the cases in the case base or training set, and similar cases are regained. The regained cases are used to advocate a solution which is reused and tested for future queries. The solution is then revised. Finally the new problem and the final solution are remembered as part of a new case.

Reusing the remembered case solution in the context of the new case is based on the idea of recognizing the variances between the remembered and the new case; and identifying the part of a remembered case which can be transported to the new case. Usually the solution of the remembered case is transported to the new case directly as a solution to this case.

Regaining the case solution generated by the reuse process is required when the solution verifies inappropriate. This offers an opening to acquire from failure.

Remembering the case is the process of integrating whatever is beneficial from the new case into the case library. This comprises of determining what information to remember and in what form to remember it, how to direct the case for forthcoming recovery; and assimilating the new case into the case archive.

The standard word for regaining is retrieving and for remembering is retaining. Reuse and Revise are used as it is in this paper.

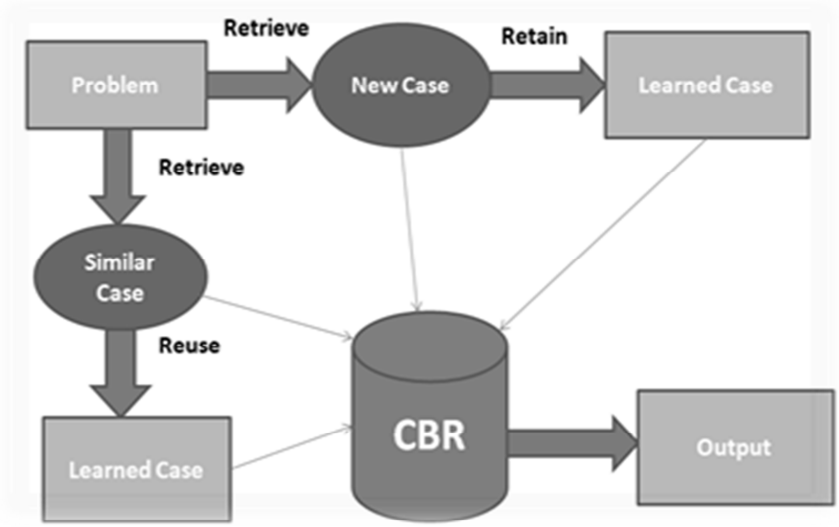

Fig.1 *Diagrammatical representation of CBR's working and the output*

3.2Steps in CBR Algorithm

a) Regain the most alike case (or cases) linking the case to the archive of past cases;
b) Reuse the regained case to try to solve the new problem;
c) Revise and acclimatize the suggested solution if required;
d) Remember the ultimate solution as portion of a new case.

3.3 Regaining a case involves

a) Recognizing a set of related problem descriptors;
b) Corresponding the case and repaying a set of satisfactorily alike cases (given a similarity count); and
c) Choosing the finest case from the set of cases refunded.

Implementation. For implementation, the values of each field; Name, Gender, Grade in A levels/intermediate, Hobbies, Skills, IQ Grade and Current Major are utilized which shape the training set. The training set is in the

form of a txt file for that matter. We have gathered around 200 cases which keep increasing because of the revise logic of the algorithm CBR.

Unique Unit Case. A separate unit case is formed from the user's entries in the system's interface, which asks the user to provide input for the same fields as stated before excluding Major, and CBR is applied. Each row of training set is compared through CBR's logic where every row of training set is compared with unit case. As each column is matched, a similarity number (3) decrements. Ultimately, the "Major" column of the row with the lowest similarity number is displayed.

3.4Algorithm for CBR Hamming Distance

Hamming Distance (count Matches: No. of attributes, sim: Similarity number, a: Training set's attributes, b: given case's attributes)

1) Check if sim=3, then, hamming distance is calculated.
2) Assign a value to count Matches according to the distance and number of attributes.
3) If the attribute of a matches with b count Match is decremented. This is done till 3rd attribute of both **a** and **b** as sim is 3.
4) If the first three attributes of **a** do not matches with **b** , then last three attributes are checked and count Matches is decremented each time it matches.
5) The difference is then calculated by dividing count Matches with 7.
6) And the major is allotted.

3.5 Algorithm for CBR Manhattan Distance

Manhattan Distance (sim: Similarity number, a: training set's attributes, b: given case's attributes)

1) Else if sim=2, then, Mnahattan distance is calculated.
2) Add up all the absolute differences of the attributes of **a** and **b**.
3) Assing the value to difference[i].
4) Update the value of difference[i] by difference[i]*6+firstAttDist/7.

Figure 2 depicts our system's functional flow chart on CBR. This figure wraps up the fore-mentioned stages in one diagram.

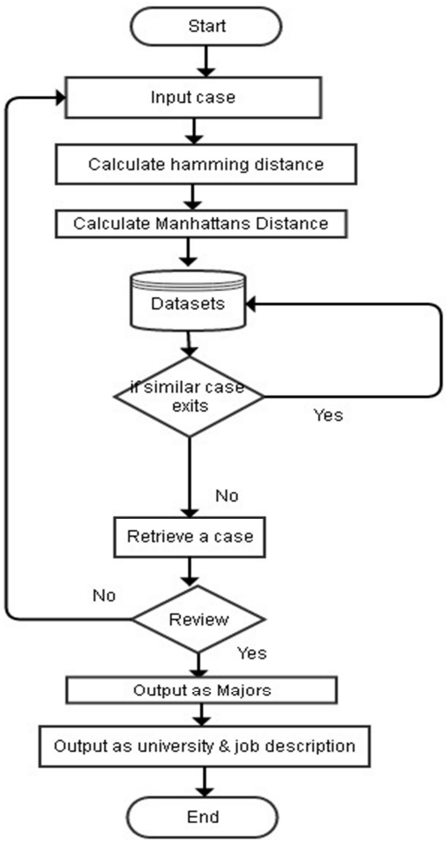

Fig.2 The Flowchart depicting the flow of our system.

4. EXPERIMENTS AND RESULTS

Data As Input. Data is selected with a keen realization of its effects on the precision of results. The data used in this paper was collected from Graduate students of 5 different Universities of Lahore, Pakistan (Table I) in the form of a questionnaire whence 20 majors' list was acquired to work on. The total count of questionnaires filled is 173. It included a set of questions that recognize the personal, educational and intelligence attributes of students of particular majors. 70% of those questions determined the IQ, and the rest were divided into questions that calculate the personality and professional concerns of the pupils. The variables attained from the questionnaire are shown in Table II.

Table I. Universities of Lahore, Pakistan: Resources and the number of questionnaires filled from each count.

Resources	Count(Datasets)
Kinnaird College	50
UCP	20
PU(Old Campus)	20
PU(New Campus)	80
CFE	3

The attributes, as shown in Table II, include general demographic information like Name and Gender. The target population was Graduate and Post Graduate students. For this questionnaire they were provided with a list of hobbies that generally interest that age group, including activities resembling outdoor sports, collecting items, playing musical instruments or listening to music, creativity based activities such as reading or writing. The students were expected to choose as many hobbies that appealed them, thus determining the attribute for Hobbies. The questionnaire also included a list of skills that are generally found in professionals working in the fields to better accommodate the user to a specific career. Grade in A levels/intermediate required the target population to fill in their academic grade in FSc./FA/ICS./I.Com to yet again better understand the academic improvements of the students studying a certain major. IQ Quiz Grade is the grade that these students attained in the quiz provided along with the questionnaire to aptly judge their intelligence and IQ tendencies to undertake a major. The quiz analyzed their command on language, mathematical knowledge, observation and problem solving techniques. In Current Major, the students provided their major which aids the system to classify the similar attributes of each major and advise careers to the users accordingly.

Table II. Variables Related To Students

Fields	Values
Name	A-Z
Gender	Male, Female
Skills	Technical, Persuasive, Entrepreneurial, etc.
Hobbies	Reading, Writing, Music, Contemplating, Relaxing, etc.
Grade in A levels/Intermediate	A, B, C, D, F

4.2 Result As Output

Majors proposed may be more than one depending on the user's input. The use case is then entered in the training set as a revised data point. After the majors are proposed, respected Universities are also proposed to the student as well as the jobs description and nature of field.

4.3 Decision Tree J 48

In this system, Decision Tree J-48 is employed through Weka, a software collection of machine learning algorithms. J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. With this technique, a tree is constructed to model the classification process for proposed majors (Fig. 1). The 173 datasets collected were input in a .txt file and then converted to arff file for Weka compatibility. Once the file is loaded in Weka and J-48 is run on it, a tree is built. This tree is applied to each tuple in the data set and results in classification for that tuple.

4.4 Algorithm for constructing Decision Tree J-48, given a data set Make-Decision-tree (DS: data set)

1. Classify the classes out of data set. (e.g.: A may be one class)
2. If (all elements of data set DS creates a class A)
3. Make a leaf node and label it as A
4. Else
5. Construct sub-trees out of DS
6. Repeat steps **1-5** until classification gets completed

4.5 Implementation in Weka

Figure.3 contains the tree view generated after running the algorithm. The squared brackets show the splitting criteria. This is the attribute name on which the parent node was split and the value (numeric) and nominal value (set) that has led to this child. The class value (Major in this case) in single quotes states the majority class in this node. The value in round brackets states (x of y) where x is the quantity of the majority class and y is the total count of examples in this node.

The first node in the figure shows majors-taken percentage according to grade and then in this grade, according to gender.

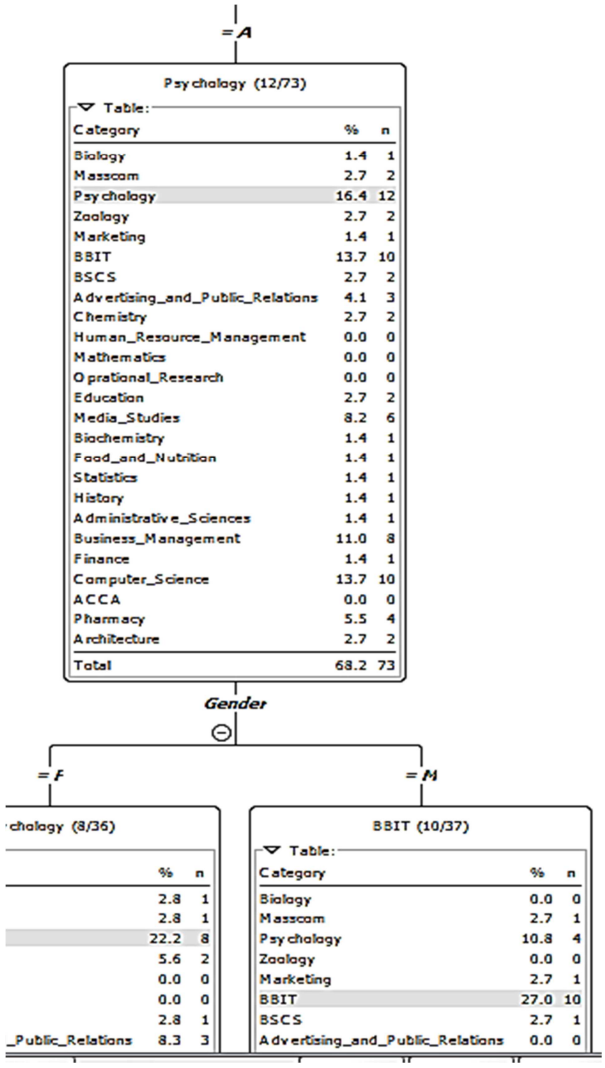

Fig.3J48 decision tree representation from Weka. The figure is only one branch of the tree depicting that each level node proceeds with the different attributes.

4.6 Comparison of Results

As CBR is an instance based algorithm, it generates different outputs by giving priority to the similarities between the test case and the collected data sets. In the light of the data sets, if the test case matches exactly 6 attributes similar to the Computer major's case, matches 5 attributes similar to the Mathematics major's case and 4 attributes with Physics major's case, then the system gives different priority on the basis of outputs. For the above mentioned scenario, the output is Computer Science, Mathematics and then Physics. Moreover, the revised concept leads to the same result again and again showing the high accuracy of the system that is approximately 80%. While comparatively using J48, the output is a definite positive decision (yes) or negative decision (no) for a student to whether go with the result or not. Only a single output is provided and accuracy of the algorithm is 50-60%, as shown in Fig 4.

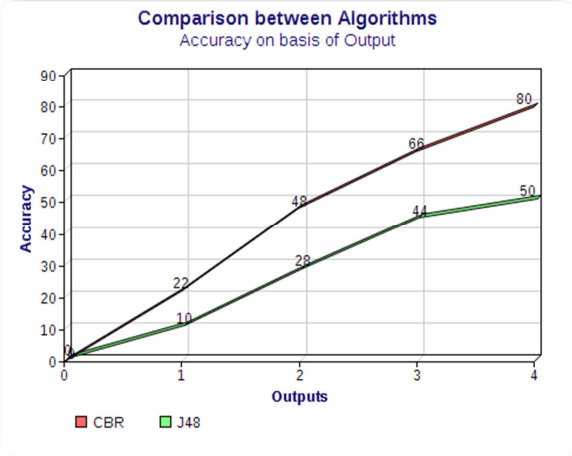

Fig. 4 Graph representing the comparison between CBR and J48

5. Limitations

The system is restricted for a certain geographical area. Culture impacts the career decisions of students widely. The system is trained over data acquired from the students of a restricted area and so it might give output that is more relatable and acceptable to students belonging to similar area and culture. The machine is trained through the datasets which are collected in the form of questionnaires from the people of only one country. Limited number of datasets is being used. The attributes in the datasets are specific.

6. Conclusion

Students nowadays face many problems when choosing a career path. They are often unsure of what may suit their interests and scope best. Many institutions are also incapable to provide students with proper guidance since they have no means to individually cater every students needs and suggest a career accordingly. Comparison based algorithms present a solution for this problem since they compare certain attributes of one case (student in this sense) with the previous cases collected (a student database). As this study indicated, CBR and Decision Tree J-48 can perfectly illuminate the way for students to select a career that exactly matches their skills and IQ tendency. The results indicate that the system is capable of correctly proposing majors with approximately 80% accuracy when presented with sufficient data and features. Out of the two algorithms tested, CBR gave the highest accuracy and Decision Tree J-48 gave the lowest accuracy.

7. Future Recommendations

Accuracy increases as the knowledge of the system increases. If the system is trained over a much larger library of cases, its accuracy will increase many folds. Gathering more types and aspects of usable data for building the library for the system will result in better accuracy, more reliability and will acquire the student's trust. The system is open-ended can certainly be extended. If the judgment criteria has some entity in it which should be cognitive of a person's psychological attribute, the system can be made more productive.

REFERENCES

1.Wong, C. (2009). Mail's Globe Careers.

2.Hall, A., (2005). Completed in Partial Fulfillment of the Requirements of PSY 8760 – Vocational Psychology.

3.Pal D., Pratap B.,&Saini,(2014).Data Mining: A Path for Effective Counseling and Course Selection.

4.Kakavand S., Mokfi T., &Jafar M. Tarokh, (2014). Prediction of the Loyal Student Using Decision Tree Algorithms.

5. Zhang, Y.,(2012).Research about the college students career counseling expert system based on agent.

6. Michael, J., & Stebleton, (2007). Career Counseling With African Immigrant College Students: Theoretical Approaches and Implications for Practice.

7. Kumar, S.,Pal, S., Nagar, J.P. Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification.

8.Kumar B., Pal, S.,(2011).Mining Educational Data to Analyze Students Performance.

9.Conati, C., Abigail, S.,Gertner, VanLehn, K., Marck, J., Druzdel, (1997). On-Line Student Modeling For Coached Problem Solving Using Bayesian Network. In *Sixth International Conference on User Modeling (UM-97).*

10.Joachim, P., Hasebrook, L., &Nathusius, W., (1997). An expert advisor for vocational guidance. *Frankfurt.*

11.Miller, H.,(2006). Building a Solution-Focused Strategy into Career Counseling. *University of Canterbury.*

12.Schedin, G.,(2007). Expectations and Experiences of Career Counseling. *Sweden.*

13. Mary, D.,&Feduccia. (2003).Career Counseling for College Students: The Influence of a Computer-Assisted Career Decision-Making Program on the Stability of College Major Selection at a Research-Extensive University.

14.Crozier, S., &Dobbs, J. Career Counseling Position Paper. *Calgary.*

15.Steven, D., Brown, &Lent,R. W., (2002). A Social Cognitive Framework for Career Choice Counseling. *Chicago.*

16.Betz, N. E., &Hackett, G. (2014). Applications of Self-Efficacy Theory to Understanding Career Choice Behavior. *Santa Barbara.*

17.Yinghua, Y. (2014).Role of Career Decision-Making Self-Efficacy and Risk of Career Options on Career Decision-Making of Chinese Graduates. *Zhejiang.*

18.Kang, R., &Lee, M. J.(2011). The Influence of Adolescent's Career Attitude, Occupation Value, and Social Support on Career Preparation Behavior. *Korea.*

19. Suling, Y. (2012). New Pattern of Dual Goal Integration of College Career Planning and Guidance.

20. Westbrook, B. (1990). The Relationship between Career Maturity Test Scores and Appropriateness of Career Choices: A replication. In *Journal of Vocational Behavior* (Vol. 36, pp. 20–32).

[21] Khan, M., & Shabbir, J. (2013). A General Class Of Estimators For Finite Population Mean In The Presence Of Non-Response When Using The Second Raw Moments. VFAST Transactions on Mathematics, 2(2), 19-36.

[22] Jan, A., & Khan, S. A. (2013). "Review of different approaches for optimal performance of multi-processors". VFAST Transactions on Software Engineering, 1(2), 7-11.